Food Control 59 (2016) 53–58

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

A modeling framework to accelerate food-borne outbreak investigations

Kun Hu*, Sondra Renly, Stefan Edlund, Matthew Davis, James Kaufman

IBM Research, Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA

ARTICLE INFO

Article history:
Received 18 April 2014
Received in revised form
4 May 2015
Accepted 12 May 2015
Available online 14 May 2015

Keywords:
Food safety
Food-borne infections
Geospatial data
Geospatial modeling
Likelihood-based method
Epidemiology
Outbreaks
Public health informatics

ABSTRACT

Food safety procedures are critical to reducing pathogen caused food-borne disease (FBD). However there is no way to completely eliminate the risk of consuming contaminated products. When prevention efforts fail, rapid identification of the contaminated product is essential. The medical and economic losses incurred grow with the duration of the outbreak. In this paper we show that before an outbreak occurs, analysis of food sales data, as a proactive intervention, can provide useful product intelligence that we can exploit during an outbreak investigation to accelerate the identification process. Using real grocery retail sales data from Germany, we have implemented a likelihood-based approach to study how such data can be used to accelerate the investigation during the early stages of an outbreak.

1. Introduction

Food-borne disease (FBD) is a global public health problem that affects millions of people every year and is caused by contamination by a variety of pathogens including bacteria, viruses, and parasites. The European Centre for Disease Prevention and Control (ECDC) gathers and reports incidence data on common pathogens that cause food-borne illness across Europe including *Norovirus*, *Campylobacter*, *Salmonella*, *Shigella*, *Listeria*, *Escherichia coli* (VTEC), and Hepatitis A (European Centre for Disease Prevention and Control, 2012). Healthcare clinicians report suspect and confirmed cases of food-borne disease to public health authorities. An outbreak is when two or more confirmed case reports are linked to the same pathogen after digesting a common food or ingredient (Rocourt, Moy, & Schlundt, 2003). Public health officials investigate outbreaks to try to identify the common food or ingredient as quickly as possible to remove the contaminated product from sale and restore consumer trust in the safety of the food supply (Marvin et al., 2009).

There have been many outbreaks where it has been difficult to identify the contaminated product using current best practices (Januszkiewicz et al., 2012; Reingold, 1998; Scavia et al., 2013). Best practices include very thorough questionnaires about food consumption for both individuals who got ill and those that did not. Public health officials can search the home for discarded containers and stored foods. Food histories can be compared to annual product consumption surveys to identify a higher correlation of particular product consumption in the ill population than in the general population (Centers for Disease Control and Prevention (CDC), 2014; Jones & Schaffner, 2003). Also, public health officials can obtain permission to retrieve data on customer loyalty program or warehouse membership card for grocery purchases (Barret et al., 2013; Gieraltowski et al., 2013). Even with these best practices, public health officials face a significant challenge and longtime delays in obtaining critical information to help identify the contaminated product.

Historically, roughly 40% of outbreaks occur from consumption from sources other than food served in restaurants or institutions (e.g., schools, nursing homes, prisons) (Jones & Angulo, 2006). Much of this food is purchased from grocery retailers. These retailers survive on razor thin profit margins and as a result have significant financial incentives to invest heavily in information

* Corresponding author.
E-mail address: khu@us.ibm.com (K. Hu).

http://dx.doi.org/10.1016/j.foodcont.2015.05.017

technology to help them efficiently manage their inventory. The grocery retailers have thus accumulated large data sets of near real-time information about food sales by stores that can provide accurate and timely information about the population's food consumption patterns. We hypothesize that before an outbreak even occurs, analysis of all retail food sales data, as a proactive intervention, can provide useful food product intelligence we can exploit during an outbreak investigation.

Kaufman et al. (2014) obtained a retail dataset that includes 580 anonymous products in two undefined food product groups from grocery retail stores in Germany. The dataset is comprised of 8176 unique retail stores distributed across 3518 postal code areas. The dataset includes the amount of each product sold per week per store during the years from 2008 to 2010 (i.e., 157 weeks in total). Leveraging this set of empirical data from Germany, we showed that before an outbreak occurs, analyzing food sales data using a proposed likelihood-based method provides useful product intelligence, and the resulting food consumption model enables one to accelerate identification of a contaminated product during the early stage of a food-borne disease outbreak. This work builds upon Kaufman et al. (2014), introducing several new and important measurements for the statistical model as well as a description of how to implement these new algorithms as components of a future system.

2. Methods

Our predictive analytics framework was created with three exchangeable components. The first component is a food distribution model that predicts where product consumption occurs. The second component is an outbreak generator. This generator uses information from the food distribution model and creates a simulated set of linked geo-coded public health case reports. The synthetic case reports capture hypothetical ill persons who consumed the contaminated product. The third component in the framework applies a statistical analysis to rank products based on the probability that each product is the cause of the outbreak. We calculate the probability leveraging the product sales distribution and location of the geo-coded public health case reports. By creating this as a flexible framework, we are able to study the individual effects of each component on performance of the system as a whole. It also enables us to compare models and methodologies, and maintain the ability to quickly load new data sets for study.

2.1. Food distribution model component

Grocery retail sales data provides us with temporal and geospatial information on food sales but not food consumption. A food distribution model is employed to show where the food is consumed in a population. Literature offers general retail shopping models such as the Huff Gravity Model (Huff, 1963), though retailers today could leverage their customer address information from loyalty or membership card programs to create a more precise spatial model.

In this work, we use a simple food distribution model $f_s(n, r)$ that assumes each product n is distributed and consumed only within a postal code region r where the product was originally purchased shown in Eq. (1). We feel this is a reasonable approach in high population density regions where people shop frequently in neighborhood markets. So if $f_c(n,r)$ is the probability that product n is consumed in region r, and $f_s(n,r)$ is the probability that product n is *sold* in region r, in our simplified model we assume:

$$f_c(n, r) = f_s(n, r) \tag{1}$$

Let $sales(n, r)$ represent the number of units of food product n sold in region r over this three-year period. We can now define a function $f_s(n, r)$ representing the probability that product n is sold in region r as:

$$f_s(n, r) = \frac{sales(n, r)}{\sum_{\widehat{r} \in R} sales(n, \widehat{r})} \tag{2}$$

The food sales data is aggregated and normalized across all German postal codes in set R such that the sum of each product is one (refer to Eq. (3)). This model simplification allows us to focus our research on examining differences in food sales distributions across the country and isolating unique patterns.

$$\sum_{r \in R} f_c(n, r) = 1 \tag{3}$$

2.2. Outbreak generator component

An outbreak generator creates individual geo-located public health case reports based on a contamination event that can include one or more products. The generator component can be used to create synthetic outbreaks or to re-create historical outbreaks for retrospective study.

In this study, we leverage knowledge about Germany and one contaminated product's distribution to generate public health case reports for a simulated food-borne disease outbreak. The normalized food consumption data (refer to Eq. (3)) from our food distribution model is input to our Monte Carlo outbreak simulation method. We generate synthetic outbreak case reports for a selected "contaminated" product x (*where we use x instead of n to indicated a single contaminated product*). Using A. J. Walker's alias method (Walker, 1977), we draw M random locations by sampling from $f_c(x, r)$ over all locations r in R. In separate trials, synthetic case report data are generated assuming each of the 580 products, in turn, as the source of contamination. We assume the products are independent so $f_c(x, r)$ also defines the probability of a case report at location r due to contaminated product x. It is true that two "products" with different local "brands" or "ids" could in fact be the same food item simply rebranded when repackaged locally. In this case, we enable noise in the generator introducing the ability to relocate small amounts of product consumption events outside the original sales postal code, which does not require the changes to the food distribution model. Conversely, a product sold on a national scale under one single brand could become contaminated at a single point of retail site (e.g., a butcher shop at a grocery store). For the purposes of this study, the simulated case reports were generated self consistently from the retail data using the assumption that the data provided to us by product id were independent. Depending upon the spatial distribution of product x, it is likely that, during one simulated outbreak of 100 cases, multiple case reports will come from a same postal code. In this work, we defined an outbreak to include between 100 and 1000 case reports and generated 100 simulated outbreaks per contaminated product for our statistical analysis.

2.3. Statistical analytical component

The statistical analytical component creates an ordered ranking of all the known products from the most likely contaminated product to the least for each synthetic outbreak generated. This component is designed to be completely independent of the food distribution model in order to create a plug and play environment that best supports fine-tuning to component-specific requirements.

A likelihood-based method (Doerr et al., 2012; Kaufman et al., 2014) is adopted in our framework to determine the probability of each product being considered as the source of illness. Let θ be a parameter vector of length N, such that the k:th element of θ is 1 if a product k is contaminated and zero otherwise. We assume there is a single contaminated product in a given outbreak so only one element of vector θ is 1. If we consider θ_k to be the parameter vector designating k as the contaminated product, then the likelihood of θ_k after observing m case reports is:

$$L(\theta_k) \propto \prod_{i=1\cdots m} f_c(k, r_i) = P_k(m) \quad (4)$$

where $f_c(k, r_i)$ is the probability that an individual living in location r_i consumed product k (refer details in the Appendix). Hence each element $P_k(m)$ of the vector $P(m)$ is proportional to the likelihood that product k is the contaminated product. Dividing each element of $P(m)$ by the largest element in $P(m)$ yields the likelihood ratio for each product being the contaminated product given the first m elements of R. We denote this as $\overline{P}(m)$. The product k that corresponds to the maximal element of $\overline{P}(m)$ is our maximum likelihood estimate for the contaminated product.

The postal codes from case reports are saved in the order produced by the generator. In first-in first-out (FIFO) order the postal codes are queried to obtain the food consumption ratios for all known products. With each new case report, the consumption ratio for each product is aggregated for the entire series of observed case reports and then ranked from highest probability to lowest probability.

2.4. Metrics

The framework currently collects four metrics in order to evaluate the effectiveness of our predictive analytic method and to characterize products in a meaningful way that provides additional insight during an outbreak investigation. For each metric collected we record the statistically robust mean (average) across a sample of 100 simulated outbreaks.

2.4.1. Success rate

The success rate defined in this study is the robust mean (average) value of simulated outbreaks where the contaminated product is correctly identified at a pre-set number of case reports (e.g., 8 case reports) across 100 simulated outbreaks. It expresses how successful the likelihood-based method is when searching for the contaminated product within our large retail data set with 1 to be the most successful case. In general, increasing the number of case reports improves our success rate (refer to Fig. 1), but our goal with this metric is to demonstrate the possibility of epidemiological value within the smallest range of case report numbers. It also helps to identify products with very low success rates regardless of how many case reports are generated.

2.4.2. First Appearance in Ordered List (FAOL)

During the earliest stages of a food-borne disease outbreak, it can be beneficial if the laboratory test on suspect foods could start earlier as triggered by a high likelihood value predicted from this method. We could provide a list of food items appearing in the top five or three most probable product list instead of focusing only on the top single one from the result of prediction. The early testing of these most suspect products can potentially save lives and costs associated with the outbreak. We calculate the robust mean (average) number and standard deviation of case reports needed across 100 simulated outbreaks where the real contaminated product (known from the assumption) first appears in the top five or three most probable ordered list (FAOL). Two examples are shown in Table 1.

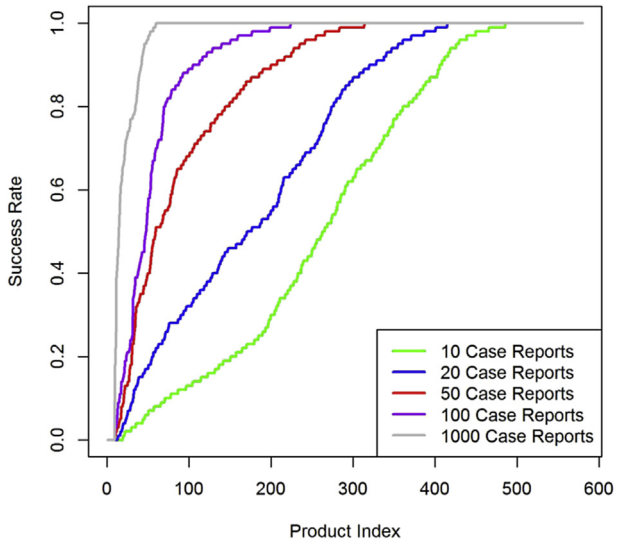

Fig. 1. Average success rate for all products at 10, 20, 50, 100 and 1000 case reports. Normally, the more case reports used in the estimation, the higher success rate we could obtain from the likelihood-based method, and more products if comtaminated can be identified. But the ultimate goal for this study is to identify the contaminated product from the list with least amount of case reports evaluated by success rate.

2.4.3. Persistent Containment in Ordered List (PCOL)

Due to the geo-location variability of the case reports, the real contaminated product is likely to move in and out of the top five or three ordered list as geo-related case reports continue to be generated. Stability occurs when the contaminated food distribution pattern and case report distribution pattern aligns. We collect the robust mean (average) number and standard deviation of case reports needed across 100 simulated outbreaks where the contaminated product remains in the top five or three most probable ordered product list. We name this metric as Persistent Containment in Ordered List (PCOL) in this study. The results are included in Table 1 along with peer FAOL metrics.

Success rate is only effective when the food remains in the top one slot of the ordered list. But PCOL provides a measure of variance between when a food is likely to remain within a top tier position in the most probable ordered list. Note that this metric is bound by the number of case reports chosen to generate in the simulated outbreak. It is possible that in a larger outbreak (e.g., beyond the 100 case reports used in this study), the contaminated product

Table 1
Mean (standard deviation) of case reports needed across 100 simulated outbreaks for two example products (assumed as real source of contamination) to first appear in Top 5, Top 3, or Top 1 of the predicted highest likely products for the metrics gathered on First Appearance in Ordered List (FAOL) and Persistent Containment in Ordered List (PCOL).

Metrics	Product A	Product B
FAOL Top 5	10.35 (6.2)	21.5 (13.99)
FAOL Top 3	13.37 (7.71)	27.3 (19.25)
FAOL Top 1	20.63 (11.33)	45.31 (29.46)
PCOL Top 5	14.44 (8.7)	28.68 (19.32)
PCOL Top 3	18.94 (10.84)	41.83 (29.55)
PCOL Top 1	30.93 (16.56)	77.92 (50.38)
Average success rate at 100 case reports	0.7303 (0.33)	0.4654 (0.29)

could be bumped out of the top tier position because of the alignment of food distribution and case report distribution. .

2.4.4. Rate of Appearance in Ordered List (RAOL)

After determining a product's success rate, we collect the average number of times each product held a position in the top three (PCOL_Top_Three), up to twice the number of case reports required to identify the contaminated product based on success rate across 100 simulated outbreaks. This metric relies on varying the number of case reports generated by product to ensure that we attain successful identification but not over-represent the products which are consistently high ranked (once the product is successfully identified). This metric, Rate of Appearance in Ordered List (RAOL), provides a measure of which products are most frequently associated with the actual contaminated product in the prediction when the framework is converging towards a steady state. These products have highly correlated distribution patterns and should all be considered during an investigation.

3. Result and analysis

Performance of the success rate metric yields an interesting result that encourages further exploration. For a fixed number of case reports (e.g., 10, 20, 50, 100, 1000), the success rates by product are shown in Fig. 1. Our method yields success rates higher than 80% for 216 products out of 580 items using as few as 10 case reports in the simulation. As expected, more case reports are used in the prediction, more products can be identified with a higher success rate. On the other hand, for 14 products out of 580 studied items, even with 1000 cases, success rate remains less than 20%, and some products never achieve a success rate above zero.

Fig. 2 shows the trajectories of improved success rate when more case reports are considered in the prediction. Two of the three (shown in green and red in Fig. 2) converge to successful identification of the contaminated product. Our dataset is well stratified in providing products whose distributions repeatedly resolve quickly and products that do not. Products that do not converge were found to have highly correlated distribution patterns with a set of products that we identified using our embedded RAOL metric and confirmed externally to our framework using well known statistical methods.

In Fig. 3, we illustrate the First Appearance in Ordered List (FAOL) metric. The products are sorted by First Appearance in the Top One Ordered List (i.e., FAOL_Top_One) in Fig. 3. This sorting creates a visual consistency in prediction when the product will appear in the Top 3 or Top 5. By measuring success when a product first appears within a slightly larger ordered list (e.g., top five), we are able to show a significant improvement. Using the FAOL Top 5 metric, 90% of outbreaks can identify the contaminated product in just 20 or fewer case reports. While this metric creates a significant theoretical improvement, its impact during a real outbreak investigation is minimized by products routinely moving in and out of the larger ordered list based on the geo-location variability of case report and the ranking of our likelihood-based estimation. The future is uncertain but can be statistically bound by the Persistent Containment in Ordered List (PCOL) metric shown in Fig. 4 and the Rate of Appearance in Ordered List (RAOL) metric shown in Fig. 5 below.

For simulations with up to 1000 case reports, we observe in Fig. 4 the bounding effect of collecting the Persistent Containment in Ordered List (PCOL) metric. The results are sorted by Persistent Containment in the Top One Ordered List (i.e., PCOL_Top_One) in Fig. 4. This metric measures the stability of the framework components in identifying the correct contaminated food product. The output of PCOL in Fig. 4 shows that the likelihood method becomes stable within 100 cases for 90% products. The method achieves better performance when we relax the constraint to be the top three and top five PCOL lists shown in red and green colors.

We show a representative plot of a cluster of highly correlated food distribution patterns recurring consistently together in the ranking of likelihood-based estimation and measured by the Rate of Appearance in Ordered List (RAOL) metric shown in Fig. 5. In this plot for contaminated product 1, we can see that products 2 and 3 occur frequently with product 1 over the duration of an outbreak. The variance in frequency establishes a pattern that could be used

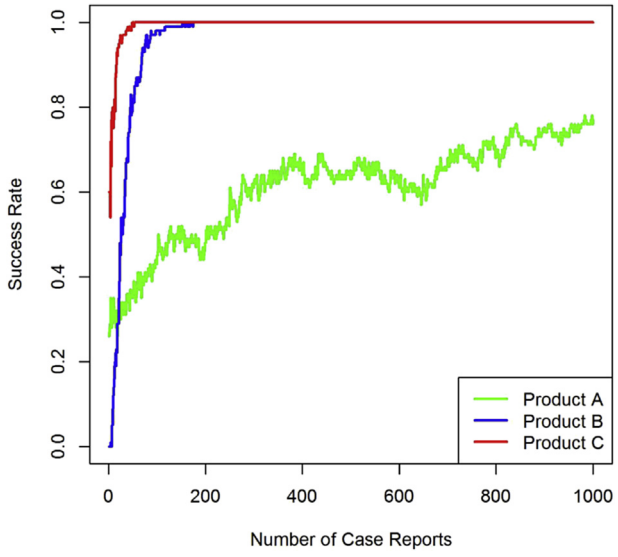

Fig. 2. The convergence of success rate for three representative product. The result shown in this graph indicate the stratification of the empirical data set used in this study. For some products, like product B (blue curve) and C (red curve), the success rates of identification if contaminated convege to 1 (the most successful) in a really fast mode when increasing the number of case reports. However, for product A (green curve), even with 1000 case reports, the trajectory of success rate does not converge to 1 at all.

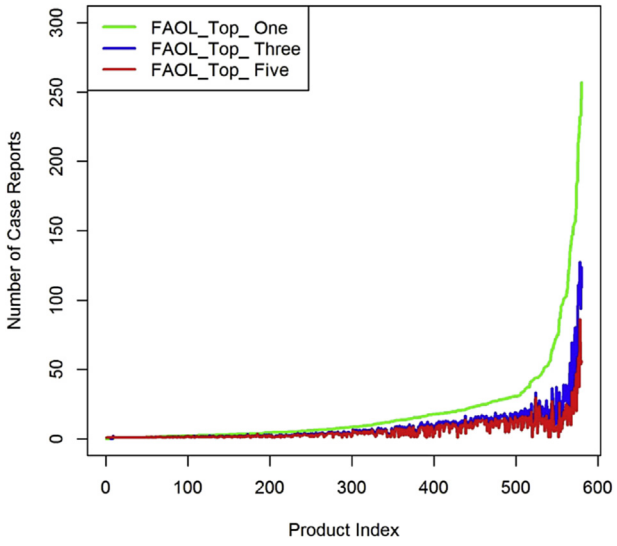

Fig. 3. First Appearance in Ordered List (FAOL). The products are sorted by First Appearance in the Top One Ordered List in this graph (blue curve). So for each product shown on X axis, it is observed that there is a consistency of identified a contaminated product in the top 3 or top 5 FAOL list. During an early stage of outbreak investigation, using the FAOL Top 5 metric, 90% of outbreaks can identify the contaminated product in just 20 or fewer case reports. This will greatly accelerate the investigation process.

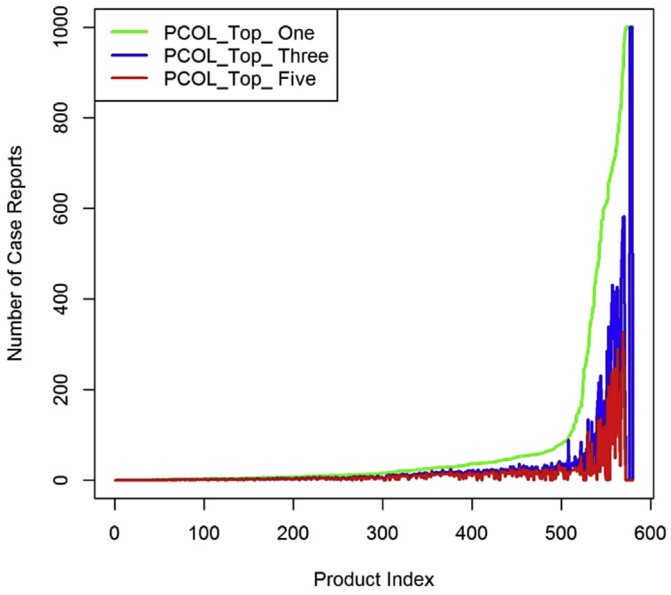

Fig. 4. Persistent Containment in Ordered List (PCOL). The data is sorted by the Persistent Containment in the Top One Ordered List (i.e., PCOL_Top_One). This metric measures the stability of the framework components in identifying the correct contaminated food product. It is observed that the likelihood method becomes stable within 100 cases for 90% products. The method achieves better performance when we relax the constraint to be the top three and top five PCOL lists shown in red and green colors.

to accelerate the identification of products with similar distributions but statistically significant differences in the RAOL metric. Identifying a group of products that are frequently observed together when the framework is converging towards a steady state suggests that the framework is nearing a convergence and could provide additional confidence during an investigation.

Following the generation of these four metrics, our analytical method creates a composite characteristic for each food product that is relevant within a given time window and for a given retail sales dataset. This composite describes the product's statistical likelihood to appear as the contaminated product at varying case report thresholds and product likelihood thresholds. It also describes the observed frequent product clustering that suggests inspection of the whole cluster. When a real food-borne disease outbreak is detected, the product composites can be exploited during the investigation to accelerate the identification process.

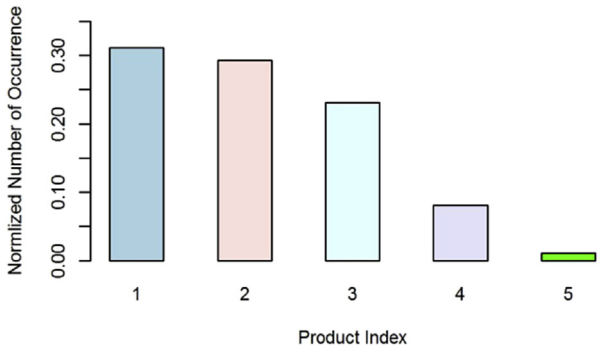

Fig. 5. Rate of Appearance in Ordered List (RAOL) for a cluster of highly correlated food products. This group of products recurs consistently together in the ranking of likelihood-based estimation method. In this plot, for contaminated product 1, we can see that products 2 and 3 occur frequently with product 1 over the duration of an outbreak. The variance in frequency establishes a pattern that could be used to accelerate the identification of products with similar distributions but statistically significant differences in the RAOL metric.

4. Discussion

This work suggests a framework that can provide useful product intelligence to accelerate the identification process during an outbreak investigation. We introduced three distinct system components in our framework; a food distribution model, an outbreak generator, and a statistical analytic component that enable us to study system performance, compare models and methodologies, and upload new data sets. Our results demonstrate the value of exploring proactive analysis to provide real-time patterns of product(s) consumption, and metrics to accelerate the investigation of food borne diseases.

Outbreak investigations have always focused on data driven analysis and best practices are continually evolving as new data sources become available, most recently with customer loyalty and membership cards program (Barret et al., 2013; Gieraltowski et al., 2013). It is a significant evolution forward when adopting the modeling approach to generate meaningful input metrics obtained by the union of terabytes of near real-time retail data with public health laboratory data obtained during an epidemiological investigation. With any modeling framework, it is critical to understand the underlying data sets, model assumptions, and statistical analysis. The conditions that influence the survival and growth of food-borne pathogens can occur at different points in the supply chain. For example, a shipment of meat might be contaminated in a slaughterhouse, thus impacting multiple retail stores and their customers, or a local butcher counter in one store could cause contamination by improper handling or storage of their local shipment. It is for this reason that the retail data should be proactively analyzed and analyzed in different ways. Meat sales by brand or shipment can be stored as one sales distribution, and sales from each and every retail store for the same ingredient can be proactively processes as separate sales distributions. In this way, comparing the outbreak data to all possible distributions, contamination at various points in the supply chain may be identified as well as the cause and effect – namely where the factors that led to growth of the pathogenic organism contributed to the outbreak.

Our analysis is performed on a well distributed set of products across Germany. The anonymous products are not unique to the stores in the dataset and are sold widely by other commercial outlets. We assume that the data sets obtained will always be incomplete, which means that we will never have all the stores or all the products. While proactive analysis of food sales data needs to be validated with real historical outbreaks, we believe that a representative data set with an appropriate food distribution model has the ability to provide seasonally sensitive and accurate intelligence about product consumed in the population. This greatly surpasses annual food surveys and provides important brand differentiation not currently available.

Our food distribution model component is additionally data dependent. Leveraging available geo-spatial information can include geographic unit proximity, transportation, population, population demographics, population socioeconomic indicators, and retailer saturation. Sophisticated models can include predicting consumer shopping behavior (e.g., (van Asselt, Meuwissen, van Asseldonk, Teeuw, & van der Fels-Klerx, 2010; Bawa & Ghosh, 1999; Bell, Ho, & Tang, 1998; Lichtenstein, Ridgway, & Netemeyer, 1993)) at the product level and take into account seasonal and local variations.

Our statistical analysis, likelihood-based estimation method, has two significant data dependencies and limitations to consider. The food distribution model can include products that are too closely correlated to demonstrate statistical significance of their probabilities. By choosing one, this gives a single product a falsely higher success rate than its closely correlated peers. Secondly,

public health case reports linked to an outbreak can be confirmed or presumptive. These products should be identified proactively before an outbreak occurs thus identifying the dependency. Incomplete knowledge is critical to making timely progress in an investigation, though at this time we have not made any accommodations for presumptive reports in this method.

Acknowledgment

We thank Matthias Filter, Chris Thoens, Annemarie Käsbohrer, and Bernd Appel from the German Federal Institute for Risk Assessment (BfR) for providing the anonymized retail sales data set and giving valuable feedback throughout this research project, and Judith Douglas for the final editorial read-through.

Appendix A

Derivation of Likelihood.

Let $\theta = \langle 0, \cdots, 0, 1, 0, \cdots, 0 \rangle$ be a vector in which the j^{th} component is one if distributed food j is contaminated:

$$\begin{aligned}
&\mathscr{L}(\theta | i \text{ is infected and } i \text{ lives in postal code } r) \\
&= P(i \text{ is infected and } i \text{ lives in postal code } r | \theta) \\
&= P(i \text{ lives in postal code } r) \cdot P(i \text{ is infected} | i \text{ lives in postal code } r, \theta) \\
&= \varphi_r \cdot \prod_j [P(i \text{ is infected} | i \text{ bought from } j, i \text{ lives in postal code } r)]^{\theta_j} \\
&= \xi \cdot \varphi_r \cdot \prod_j [P(i \text{ bought from } j | i \text{ lives in postal code } r)]^{\theta_j}
\end{aligned}$$

where ϕ_r denotes the population density in postal code r. An individual can be a determined carrier of infection with certainty ξ. Each distributed food product j is associated with a retail store.

$$\mathscr{L}(\theta | i \text{ is infected and } i \text{ lives at } r) = \xi \cdot \varphi_r \cdot \prod_j \left[\sum_{k \in R_j} P(i \text{ bought at } k | i \text{ lives in postal code } r) \right]^{\theta_j}$$

Then having observed a set of reported cases D, the likelihood becomes:

$$L(\theta | D) = \prod_{i \in D} \left\{ \xi \cdot \varphi_r \cdot \prod_j \left[\sum_{k \in R_j} P(i \text{bought at} k | i \text{lives in postal code} r) \right]^{\theta_j} \right\}$$

Assuming $f(i,k)$ denotes the probability that customer i shops at store k given i lives at x,y, we arrive at the following objective function:

$$O(\theta | D) = \sum_{i \in D} \left\{ \log \xi + \log \varphi_r + \sum_j \theta_j \log \left[\sum_{k \in R_j} f(i,k) \right] \right\}$$

Since the first two terms are constant, we can simplify to:

$$O(\theta | D) = \sum_{i \in D} \sum_j \theta_j \log \sum_{k \in R_j} f(i,k)$$

We maximize this function to determine the most likely contaminated food.

References

van Asselt, Meuwissen, M., van Asseldonk, Teeuw, J., & van der Fels-Klerx, H. J. (2010). Selection of critical factors for identifying emerging food safety risks in dynamic food production chains. *Food Control, 21*(6), 919–926.

Barret, A. S., Charron, M., Mariani-Kurkdjian, P., Gouali, M., Loukiadis, E., Poignet-Leroux, B., et al. (2013). Shopper cards data and storage practices for the investigation of an outbreak of Shiga-toxin producing *Escherichia coli* O157 infections. *Médecine et Maladies Infectieuses, 43*(9), 368–373.

Bawa, K., & Ghosh, A. (1999). A model of household grocery shopping behavior. *Marketing Letters, 10*(2), 149–160.

Bell, D. R., Ho, T. H., & Tang, C. S. (1998). Determining where to shop: fixed and variable costs of shopping. *Journal of Marketing Research*, 352–369.

Centers for Disease Control and Prevention (CDC). (2014). *Foodborne active surveillance network (FoodNet) population survey atlas of exposures* (pp. 2006–2007). Atlanta, Georgia: U.S.: Department of Health and Human Services, Centers for Disease Control and Prevention.

Doerr, D., Hu, K., Renly, S., Edlund, S., Davis, M., Kaufman, J. H., et al. (2012). Accelerating investigation of food-borne disease outbreaks using pro-active geospatial modeling of food supply chains. In *Proceedings of the first ACM SIGSPATIAL international workshop on use of GIS in public health* (pp. 44–47). Redondo Beach, California: ACM.

European Centre for Disease Prevention and Control. (2012). *Annual epidemiological report. Reporting on 2010 surveillance data and 2011 epidemic intelligence data*.

Gieraltowski, L., Julian, E., Pringle, J., Macdonald, K., Quilliam, D., Marsden-Haug, N., et al. (2013). Nationwide outbreak of *Salmonella* Montevideo infections associated with contaminated imported black and red pepper: warehouse membership cards provide critical clues to identify the source. *Epidemiology and Infection, 141*(6), 1244–1252.

Huff, D. L. (1963). A probabilistic analysis of shopping center trade areas. *Land Economics, 39*(1), 81–90.

Januszkiewicz, A., Szych, J., Rastawicki, W., Wolkowicz, T., Chrost, A., Leszczynska, B., et al. (2012). Molecular epidemiology of a household outbreak of Shiga-toxin-producing *Escherichia coli* in Poland due to secondary transmission of STEC O104: H4 from Germany. *Journal of Medical Microbiology, 61*(4), 552–558.

Jones, T. F., & Angulo, F. J. (2006). Eating in restaurants: a risk factor for foodborne diease? *Clinical Infectious Diseases, 43*(10), 1324–1328.

Jones, T. F., & Schaffner, W. (2003). *Salmonella* in imported mangos: shoeleather and contemporary epidemiologic techniques together meet the Challenge. *Clinical Infectious Diseases, 37*(12), 1591.

Kaufman, J., Lessler, J., Harry, A., Edlund, S., Hu, K., Douglas, J., et al. (2014). A likelihood-based approach to identifying contaminated food products using sales data: performance and challenges. *PLoS Computational Biology, 10*(7), e1003692.

Lichtenstein, D. R., Ridgway, N. M., & Netemeyer, R. G. (1993). Price perceptions and consumer shopping behavior: a field study. *Journal of Marketing Research, 30*(2), 234–245.

Marvin, H. J. P., Kleter, G. A., Frewer, L. J., Cope, S., Wentholt, M. T. A., & Rowe, G. (2009). A working procedure for identifying emerging food safety issues at an early stage: implications for European and international risk management practices. *Food Control, 20*(4), 345–356.

Reingold, A. L. (1998). Outbreak investigations–a perspective. *Emerg Infect Dis, 4*(1), 21–27.

Rocourt, J. G., Moy, K. V., & Schlundt, J. (2003). *The present state of foodborne disease in OECD countries*. WHO.

Scavia, G., Ciaravino, G., Luzzi, I., Lenglet, A., Ricci, A., Barco, L., et al. (2013). A multistate epidemic outbreak of *Salmonella* goldcoast infection in humans, June 2009 to March 2010: the investigation in Italy. *Eurosurveillance, 18*(11), 20424.

Walker, A. J. (1977). An efficient method for generating discrete random variables with general distributions. *ACM Transactions on Mathematical Software, 3*(3), 253–256.